ELSEVIER Computational Statistics & Data Analysis 39 (2002) 529–554

COMPUTATIONAL STATISTICS & DATA ANALYSIS

www.elsevier.com/locate/csda

Estimation of upper quantiles under model and parameter uncertainty

Reza Modarres[1,*], Tapan K. Nayak[1], Joseph L. Gastwirth[2]

Department of Statistics, The George Washington University, 2201 G Street, N.W. Washington DC, 20052, USA

Received 1 October 2000; received in revised form 1 August 2001

Abstract

In this paper we assess accuracy of some commonly used estimators of upper quantiles of a right skewed distribution under both parameter and model uncertainty. In particular, for each of log-normal, log-logistic, and log-double exponential distributions, we study the bias and mean squared error of the maximum likelihood estimator (MLE) of the upper quantiles under both the correct and incorrect model specifications. We also consider two data dependent or adaptive estimators. The first (tail-exponential) is based on fitting an exponential distribution to the highest 10–20 percent of the data. The second selects the best fitting likelihood-based model and uses the MLE obtained from that model. The simulation results provide some practical guidance concerning the estimation of the upper quantiles when one is uncertain about the underlying model. We found that the consequences of assuming log-normality when the true distribution is log-logistic or log-double exponential are not severe in moderate sample sizes. For extreme quantiles, no estimator was reliable in small samples. For large sample sizes the selection estimator performs fairly well. For small sample sizes the tail-exponential method is a good alternative. Presenting it and the MLE for the log-normal enables one to assess the potential effects of model uncertainty.

Keywords: Quantile estimation; Model uncertainty; Parameter uncertainty; Model selection; Likelihood; Tail-exponential; Log-symmetric; Monte Carlo simulation

* Corresponding author. Tel.: +1-202-994-6359; fax: +1-202-994-6917.

E-mail addresses: reza@gwu.edu (R. Modarres), tapan@gwu.edu (T.K. Nayak), jlgast@.gwu.edu (J.L. Gastwirth).

[1] Research was supported in part by a cooperative agreement with the Office of Water of the U.S. Environmental Protection Agency.

[2] Research was supported in part by a grant from the NSF and was completed while this author was a Visiting Scientist at the Division of Cancer Epidemiology and Genetics of the National Cancer Institute.

0167-9473/02/$ - see front matter
PII: S0167-9473(01)00094-9

1. Introduction

Estimation of upper quantiles of distributions is important in many applications. Estimates of the upper quantiles of the distribution of a risk factor or an exposure index are commonly used to assess the risk to human health as a result of exposure to chemicals and microbes in the environment, or to determine if concentration levels of contaminants exceed specified limits. Quantitative risk assessment using Monte Carlo methods requires selection of appropriate probability distributions for the risk factors. When the distribution of an important factor is modeled by an incorrect distribution, inaccurate risk estimates may result. Haas (1997) discussed the importance of the distributional form when specifying inputs to Monte Carlo risk assessment. In particular, he showed that the tail behavior of distributions from different families with the same mean and same variance may differ substantially when the variance is large. He also demonstrated that correctly identifying the true model with high probability requires large sample sizes. These considerations indicate that quantile estimates may also be sensitive to the assumed distributional form. Thus, it is important to examine the effects of model selection, and mis-specification on risk estimates. In this paper we investigate the accuracy and robustness of certain quantile estimators under both correct and incorrect model specifications.

The Safe Drinking Water Act requires the United State Environmental Protection Agency (USEPA) to set drinking water standards to control the level of contaminants in drinking water. The National Primary Drinking Water Regulations codify these enforceable standards. Such standards protect the public from the effects of contaminants by limiting their levels in drinking water. Maximum Contaminant Level is the highest level of a contaminant that USEPA allows in drinking water. The maximum contaminant levels for a host of microorganisms, disinfectants and disinfection by-products, inorganic and organic chemicals, and radio nuclides have been established (USEPA, 2001).

The size of the environmental samples are usually small. Sampling for compliance purposes, for example, may be required to be performed monthly or annually since the sampling process may cause disruption in the plant operation. In some instances measuring procedures and laboratory determinations for such substances in water, air, or soil samples are expensive, leading to small-sized samples. For example, Frey and Burmaster (1999) study estimates of the 95th percentile based on datasets of sizes 19, 9 and 5. Estimating upper quantiles based on small-to-moderate sample sizes cannot be avoided as it may be mandated by regulation. For example, USEPA (1985) provides guidance for setting and monitoring aquatic standards on toxic chemicals based on the estimates of the 95th percentile. The upper quantiles are often used in regulatory settings to reflect a degree of prudence in the decision-making process. This is especially true when the issues involved concern human health or the protection of natural resources.

In this paper, we focus on continuous distributions. Let the probability and cumulative distribution functions of Y be denoted by f and F, respectively. Then, the pth quantile of Y is defined as $Q_Y(p)=F^{-1}(p)$, which is the smallest y such that $F(y)=p$. Since regulatory decisions rely on the upper tail of a distribution, we are concerned

with estimating the upper quantiles and extreme upper quantiles of Y based on a random sample $y_1,\ldots,y_n$ of observations from the distribution of Y. We define upper quantiles to refer to quantiles that correspond to $0.90 \leqslant p < 0.99$ and extreme upper quantiles will refer to quantiles that are at or above the 99th percentile.

In a parametric approach, one assumes that the true distribution of Y belongs to a family of distributions $\mathscr{F} = \{f_\theta(y),\ \theta \in \Theta\}$ indexed by a parameter θ, which may be vector-valued. Under such a model, the quantiles are functions of the parameters of the model, and hence estimation of quantiles amounts to estimation of certain parametric functions, namely, estimation of $Q_Y(p) = F_\theta^{-1}(p)$, where $F_\theta(p)$ is the cumulative distribution function (CDF) of Y and θ is the true value of the parameter. These quantities can be estimated using the maximum likelihood method according to which the estimate of $Q_Y(p)$ is $\hat{Q}(p) = F_{\hat{\theta}}^{-1}(p)$, where $\hat{\theta}$ is the MLE of θ under the specified model $\mathscr{F}$. It may be noted that estimation of upper and extreme quantiles is a difficult task. Letting $p = F(y) = \int_{-\infty}^{y} f(t)\,\mathrm{d}t$, and $y = F^{-1}(p) = Q_Y(p)$, it is seen that $\mathrm{d}y/\mathrm{d}p = \mathrm{d}Q_Y(p)/\mathrm{d}p = 1/f(y)$. Therefore, the quantile values, $Q_Y(p)$, change very rapidly with p when $f(y)$ is small, i.e., in the upper and extreme upper tails of the underlying distribution. Thus, for accurate estimation of upper and extreme quantiles one needs very accurate estimate of the upper tail of the distribution. This is difficult as very few observations from the upper and extreme upper tails occur in modest sized samples.

Usually, a parametric distributional model is chosen based on physical or biological grounds (Kapteyn, 1903). For example, Ott (1995, Chapter 8) discusses the dilution of pollutants in the environment and argues that repeated dilution of a contaminant with water results in a gamma distribution. The log-normal distribution arises as the product of many independent random factors (Aitchison and Brown, 1973). In such cases, there is little uncertainty about the underlying distribution, and a suitable model can be identified a priori.

In many environmental applications, however, we do not have adequate physical, biological, or empirical knowledge to suggest the functional form of the underlying distribution, i.e., to suggest one distributional model. But, we may have enough knowledge to suggest that the true distribution belongs to one of certain specific families. In such cases, one may assume that F is a member of a set of parametric families $\mathscr{F}_i$, $i = 1,\ldots,k$ as in robustness studies (Gastwirth, 1966; Andrews et al., 1972). For example, for unimodal and right skewed variables with unbounded support, one may assume that the true distribution is either log-normal, or log-logistic, or log-double exponential. Then, one may investigate the data by various exploratory techniques (Hoaglin et al., 1983) and confirmatory tests of hypotheses (D'Agostino and Stephens, 1986) in order to identify a single model that best describes the observations. That is, the information provided by the sample may be used to identify a single best model (Draper, 1995) as measured by some criteria. The investigation can be exploratory or confirmatory in nature. There are several techniques of model selection, including optimal and sub-optimal invariant rules (Quesenberry and Kent, 1982), maximum likelihood rules (Dumonceaux and Antle, 1973; Kappenman, 1982), and rules based on goodness-of-fit statistics (Dyer, 1973). Clearly, the choice of the methods and criteria for data-based model selection involves some judgment.

Alternatively, in the presence of model uncertainty one may use non-parametric estimators, which are based on minimal and mild assumptions regarding F, such as continuity or existence of moments (Lehman, 1983). This avoids parametric model selection which requires additional assumptions about the functional form of F. In non-parametric approaches, the empirical distribution function, in various interpolated forms, or a quasi-empirical distribution are used to estimate F. The estimated distribution function is then inverted to obtain quantile estimates. Specifically, we shall consider quantile estimates that are based on the empirical quantile function, or the tail-exponential method described in Section 3.

Generally, non-parametric estimators are less efficient than parametric estimators when the assumed parametric model is correct. Parametric estimators are more attractive when there is not much uncertainty about the model. They run the risk of being inaccurate as the assumed model may not be the true model. However, the choice of the model may or may not have crucial effects on the final inferences. For example, in the context of general linear models, McCullagh and Nelder (1989) and Aitchison (1982) have suggested that assuming a log-normal distribution will in many cases produce the same conclusions as assuming a gamma distribution. On the other hand, Wiens (1999) showed that the two competing models, log-normal and gamma, yielded different conclusions in the analysis of the effects of an investigational vaccine. For estimation in regression models with multiplicative errors, Firth (1988) showed that maximum likelihood estimates based on gamma errors are more efficient than those based on a log-normal distribution under reciprocal misspecification.

To examine the effects of model uncertainty, in this paper, we consider the log-normal, log-logistic, and log-double exponential families, and investigate the properties of the maximum likelihood estimators (MLE) of certain quantiles under both correct and incorrect model specifications. For example, when the true distribution is log-normal, we investigate the bias and mean squared error (MSE) of the maximum likelihood estimators of quantiles derived under the assumption that the true distribution is log-normal (i.e., under correct model specification), as well as under the assumptions that the true distribution is log-logistic, and log-double exponential, respectively (i.e., under incorrect model specifications). We focus on maximum likelihood estimators because they are used frequently in practice. While some asymptotic properties of maximum likelihood estimators of quantiles under a mis-specified general linear model have been discussed by Séménou (1996), we investigate small sample properties by simulation. We also compare the MLEs with non-parametric estimators, and a natural estimator. For the last estimator we first select a model, among log-normal, log-logistic, and log-double exponential, and then calculate the MLE based on the selected model. In this context, we select the model for which the maximized likelihood is the largest. Thus, our selection estimator is also the MLE under the union of the three models. As the selection estimator deals with both model and parameter uncertainty, it is expected to have larger sampling variation than the MLE based on the correct model. Note that if the correct model is identified a priori, only uncertainty about the parameters leads to error in estimation of $Q_Y(p)$. Effects of model mis-specification on the width and coverage probability of confidence intervals is also an important issue, but we do not investigate it in this paper.

In the next section, we review some basic properties of the log-normal, log-logistic, and log-double exponential families of distributions. These three families are log-symmetric, and have been found useful for modeling environmental data. Also, they are location-scale families on the log scale. To compare the three families we study their quantile plots simultaneously. Section 3 considers the empirical distribution function and discusses the tail-exponential method for estimating the upper values of the quantile function. Section 4 describes a simulation study and compares the performance of the quantile estimators. An example is discussed in Section 5. The final section is devoted to summary and recommendations.

2. Distributions

In this section, we discuss three symmetric location-scale distribution families on the log scale. A location-scale family is obtained by considering location and scale transformations of a random variable with a specified distribution. Let Z be a random variable with density f_Z, distribution function F_Z, and quantile function $Q_Z(p)$. Consider the transformation $X = a + bZ$, $a \in R$, $b > 0$. Then, it can be seen that X has density $f_X(x) = 1/bf_Z((x-a)/b)$, distribution function $F_X(x) = F_Z((x-a)/b)$, and quantile function $Q_X(p) = a + bQ_Z(p)$. Further if the distribution of Z is symmetric about zero, i.e., $f_Z(-z) = f_Z(z)$ for all z, then the distribution of X is symmetric about a, i.e., $f_X(a-t) = f_X(a+t)$ for all t. If Z is symmetric about 0, it also follows that $F_Z(z) = 1 - F_Z(-z)$ for all z, and $Q_Z(p) = -Q_Z(1-p)$ for all p. The parameters a and b of a location-scale family determine the center, and the dispersion of the distribution, respectively. The normal, logistic, and double exponential are three well-known symmetric location-scale families of distribution.

We shall however, assume that the risk factor Y is such that the distribution of its logarithm belongs to a symmetric location-scale family. Thus, $Y = \exp(X) = \exp(a+bZ)$ for some $a \in R$ and $b > 0$, where Z has a known distribution which is symmetric around 0. Being a continuous and increasing function, the exponentiation leads to distributions that are unimodal and right skewed. This transformation alters the spacing while preserving the order of the observations. It can be seen that Y has density $f_Y(y) = 1/yf_X(\ln y) = 1/byf_Z((\ln(y)-a)/b)$, and quantile function $Q_Y(p) = \exp(Q_X(p)) = \exp(a + bQ_Z(p))$. Thus, the distribution, the quantile function and other properties of the distribution of Y can be obtained readily from the distribution of Z. We next review some specific properties of three distribution families. As the interpretations of the parameters a and b are different for the three families, we express them in terms of the mean μ_y and coefficient of variation, $v_y = \sqrt{Var(Y)}/E(Y)$ of Y. It may be noted that under the assumption that log Y has a symmetric location-scale distribution, the CV of Y depends only on the scale parameter b and not on the location parameter a.

2.1. Log-normal distribution

The log-normal distribution is basic in the modeling of environmental, economic, and industrial observations. It has been used to fit air quality data (Mage, 1981),

Table 1
Log-normal properties

Distribution	Log-normal
PDF, $f(y)$	$y = \exp(x), \quad x \sim N(a, b^2),$ $\frac{1}{b\sqrt{2\pi}y}\exp(\frac{-(\ln y - a)^2}{2b^2}), \quad y > 0, b > 0$
CDF, $F(y)$	$\Phi(\frac{\ln(y)-a}{b})$, Φ is the standard normal CDF
Quantile function, $Q_Y(p)$	$\exp(a + b\Phi^{-1}(p))$
Mean, μ_y	$\exp(a + b^2/2)$
Variance, σ_y^2	$\exp(2a + b^2)(\exp(b^2) - 1)$
Skewness, $\eta_3 = \mu_3/b^3$	$(\exp(b^2) + 2)\sqrt{\exp(b^2) - 1}$
Coefficient of variation, v	$\sqrt{\exp(b^2) - 1}$
log(likelihood)	$-n(\ln b\sqrt{2\pi}) - \sum_{i=1}^{n} \ln y_i - 1/2b^2 \sum_{i=1}^{n} (\ln y_i - a)^2$
Parameter MLE	$\hat{a} = 1/n \ln y_i$ and $\hat{b}^2 = 1/n(\sum_{i=1}^{n} \ln^2 y_i - \hat{a}^2)$

water consumption rates (Rosebury and Burmaster, 1992), and trace elements in human tissue (Rustagi, 1964). Many exposure factors such as body weight as a function of age (Burmaster and Crouch, 1997), total skin area as a function of body weight (Burmaster, 1998), and fish consumption rates (Murray and Burmaster, 1994) have also been modeled by log-normal distributions.

A random variable Y has a log-normal distribution if $Y = \exp(X)$, and X has a normal distribution. Here the parameters a and b are the mean μ and standard deviation σ of X. Standard calculations show that $\mu_y = \exp(\mu + \sigma^2/2)$, and $v_y = \sqrt{\exp(\sigma^2) - 1}$ (see Table 1). So, for given v_y and μ_y, the corresponding values of μ and σ^2 are $\mu = \ln(\mu_y) - \frac{1}{2}\ln(v_y^2 + 1)$, and $\sigma^2 = \ln(v_y^2 + 1)$. The quantiles of log-normal distributions are obtained from the quantiles of the standard normal distribution by noting that $Q_Y(p) = \exp(\mu + \sigma Q_z(p))$, where $Q_z(p) = \Phi^{-1}(p)$, and $\Phi^{-1}(p)$ is the quantile function of the standard normal distribution. This indicates that the upper quantiles of the log-normal distribution are affected by the scale (σ) of the underlying normal distribution. For further reviews of this distribution we refer to Aitchison and Brown (1973), and Johnson et al. (1995).

2.2. Log-logistic distribution

The logistic distribution with location parameter a and scale parameter b has the density function

$$f(x) = \frac{\exp[-(x-a)/b]}{b[1 + \exp(-(x-a)/b)]^2}, \quad -\infty < x < \infty, \quad b > 0, \quad -\infty < a < \infty.$$

One can show $E(X) = a$ and $Var(X) = b^2\pi^2/3$ (see Table 2). The distribution of $Y = \exp(X)$ is given by

$$g(y) = \frac{\exp(a/b)y^{-(1/b)-1}}{b[1 + \exp(a/b)y^{-1/b}]^2}, \qquad y > 0.$$

Table 2
Log-logistic properties

Distribution	Log-logistic
PDF, $f(y)$	$y = \exp(x), \quad x \sim L(a,b), \ b > 0$ $(1/b)\exp(a/b)y^{-1/b-1}(1+\exp(a/b)y^{-1/b})^{-2}, \quad y > 0$
CDF, $F(y)$	$(1+\exp(a/b)y^{1/b})^{-1}$
Quantile function, $Q_y(p)$	$\exp(a)(\frac{p}{1-p})^b$
Mean, μ_y	$b\pi\exp(a)\csc(b\pi)$
Variance, σ_y^2	$b\pi\exp(2a)(\tan(b\pi)-b\pi)\csc^2(b\pi)$
Skewness, $\eta_3 = \mu_3/\sigma_y^3$	*Note*: $E(y^r) = (b\pi r)\exp(ar)\csc(b\pi r), \quad r < 1/b$
Coefficient of variation, v	$\sqrt{1/(b\pi)\tan b\pi - 1}$
log(likelihood)	$n(\ln(1/b)+a/b)-(1/b+1)\sum_{i=1}^{n}\ln y_i$ $-2\sum_{i=1}^{n}\ln(1+\exp(a/b)y_i^{-1/b})$
Parameter MLE	No closed form expression

It can be shown that $E(Y^r)=(\pi rb)\exp(ra)\csc(\pi rb)$, and $v_y=\sqrt{1/\pi b\tan(\pi b)-1}$. Using these relationships we can find the values of a and b corresponding to given values of μ_y and v_y. For that we need to solve $v_y^2+1=(1/\pi b)\tan(\pi b)$ for b and calculate $a=\ln([\mu_y/\pi b]\sin(\pi b))$. For further discussion of the log-logistic distribution we refer to Johnson et al. (1995).

The log-logistic distribution has been used to model survival data (Bennett, 1983) and business failure data (Dubey, 1966). The shape of the log-logistic distribution is similar to a log-normal distribution as the normal and logistic distributions are very similar in shape. Johnson et al. (1995) show that $|[1+\exp(-\pi x/\sqrt{3})]^{-1}-\Phi(16x/15)|<0.01$, where $[1+\exp(-\pi x/\sqrt{3})]^{-1}$, and $\Phi(x)$ are the distribution functions of standard logistic and standard normal distributions, respectively. They also suggest that, on suitable occasions, the normal can replace the logistic to simplify the analysis. Due to their similarity, statisticians often do not concern themselves with whether the normal or logistic distributions underlie the data. It is also very difficult to distinguish between these two distributions at small sample sizes. Even though logistic provides a good approximation in the central part of the normal distribution, there can be substantial differences in the upper and extreme upper quantiles. With a kurtosis of 4.2, the standard logistic distribution has a longer tail than the normal, which has kurtosis 3.0. In fact the logistic distribution has been shown to be better approximated by a t distribution with nine degrees of freedom (Mudholkar and George, 1978). Differences in the upper quantiles of normal and logistic are further magnified when they are exponentiated to get log-normal and log-logistic distributions.

2.3. Log-double exponential distribution

The log-double exponential distribution is briefly discussed in Johnson et al. (1995). Let X have a double exponential distribution with location parameter a and scale parameter b. That is

$$f(x) = \frac{1}{2b}\exp(-|x-a|/b), \quad -\infty < x < \infty, \quad b > 0, \quad -\infty < a < \infty.$$

The density function of $Y = \exp(X)$ can be expressed as

$$g(y) = \begin{cases} \frac{1}{2b}\exp(-a/b)y^{1/b-1} & \text{if } 0 \leqslant y \leqslant \exp(a), \\ \frac{1}{2b}\exp(a/b)y^{-1/b-1} & \text{if } y \geqslant \exp(a). \end{cases}$$

It is straight forward to verify that $E(y^r) = [1-(rb)^2]^{-1}\exp(ra)$ for $r < 1/b$, and $1 + v_y^2 = (\alpha^2 - 2\alpha + 1)/(\alpha^2 - 4\alpha)$, where $\alpha = 1/b^2$. Then, letting $k = 1 + v_y^2$ we see that

$$(1-k)\alpha^2 + (4k-2)\alpha + 1 = 0$$

and the values of a, and b corresponding to given values of μ_y and v_y are

$$b = \left[\frac{2 - 4k - \sqrt{(4k-2)^2 - 4(1-k)}}{2(1-k)}\right]^{-1/2}$$

and $a = \ln(\mu_y(1-b^2))$ (see Table 3).

Johnson (1949) discussed a system of three transformations, including a log transformation, of normal and double exponential distributions. Tadikamalla and Johnson (1982) discussed the same system of transformations applied to the logistic distribution. Log-double exponential distribution, also called log-Laplace distribution, has been used to model dose-response data (Uppuluri, 1980).

To compare the three families of distributions Figs. 1–6 simultaneously plot their quantile functions in the upper region ($0.90 < p < 0.999$) for several identical values of the mean and CV. A change in the mean results in a shift in the plot but a change in the CV changes the shapes of the graphs and the ordering of the extreme quantiles. Examination of the three quantile functions $Q(p)$ over the entire region of p indicate that they are very close to each other over an interval $(0, p_1)$ of values of p. The value of p_1 depends on the CV and increases with the CV. For example, investigating the entire range of p and several values of CV indicate that, the value of p_1 is around 0.50 for CV = 0.1, and around 0.93 for CV = 10. As the value of p exceeds this value p_1, the quantile functions separate from each other. For small-to-moderate values of the CV, the quantile functions cross each other again at a value p_2, which is between p_1 and 1. The value of p_2 also increases with the CV. For example, p_2 is around 0.955 when the CV = 0.1, and around 0.998 for CV = 1. For much larger CV values this crossing does not occur.

Figs. 1–4 show the differences among the upper and extreme quantiles (i.e., for p between 0.90 and 0.999) of the log-normal, log-logistic, and log-double exponential distributions for $\mu_y = 1$, and CV = 0.1, 1. For the values of p that are between p_1

Table 3
Log-double exponential properties

Distribution	Log-double exponential
PDF, $f(y)$	$y = \exp(x), \quad x \sim DE(a,b)$ $\frac{1}{2b}\exp(-a/b)y^{1/b-1}, \quad 0 \leqslant y \leqslant \exp(a)$ $\frac{1}{2b}\exp(a/b)y^{-1/b-1}, \quad y \geqslant \exp(a)$
CDF, $F(y)$	$\frac{1}{2}\exp(-a/b)y^{1/b}, \quad y \geqslant \exp(a)$ $1 - \frac{1}{2}\exp(a/b)y^{-1/b}, \quad y \geqslant \exp(a)$
Quantile function, $Q_Y(p)$	$(2p\exp(a/b))^b, \quad 0 \leqslant p \leqslant 1/2$ $(2(1-p)\exp(-a/b))^{-b}, \quad 1/2 \leqslant p \leqslant 1$
Mean, μ_y	$(1-b^2)^{-1}\exp(a), \quad b < 1$
Variance, σ_y^2	$\exp(2a)((1-4b^2)^{-1} - (1-b^2)^{-2}), \quad b < 2$
Skewness, $\eta_3 = \mu_3/\sigma_y^3$	*Note*: $E(y^r) = [(1-(rb)^2]^{-1}\exp(ra), \quad b < 1/r$
Coefficient of variation, v	$\sqrt{\frac{(1-b^2)^2}{1-4b^2} - 1}, \quad b < 2$
log(likelihood)	$-\ln b - a/b + (1/b-1)\ln y_i - \ln 2, \quad 0 \leqslant y \leqslant \exp(a)$ $-\ln b + a/b - (1/b+1)\ln y_i - \ln 2, \quad y \geqslant \exp(a)$
Parameter MLE	$\hat{b} = \frac{1}{n}\sum_{i=0}^{n} \lvert x_i - \hat{a}\rvert$ $\hat{a} = median(x_i)$

and p_2, the quantiles for the log-normal are larger than the corresponding quantiles under log-logistic, which in turn are larger than those for the log-double exponential. This ordering is reversed for values of p that are larger than p_2 when the crossing takes place, i.e., for small values of the CV. Thus, for the same mean and CV, the ordering of the quantiles $Q(p)$ under the three models depends on p as well as the CV. It is interesting to note that for small and moderate CV the log-double exponential has the longest tail, and log-normal has the shortest tail but, this ordering is reversed for large CV.

3. Empirical distribution function

Not knowing the general form of F, we would ideally like to find accurate estimates of $Q_Y(p) = F^{-1}(y)$ for $0 < p < 1$ under mild assumptions, such as existence of moments, or continuity of F. In such settings, $F_n(y) = 1/n\#\{y_i \leqslant y\}$, the empirical distribution function, is a natural estimator of $F(y)$. This estimator places equal probability mass $1/n$ at each sample point y_i and is the non-parametric maximum likelihood estimate of $F(y)$ (Efron and Tibshirani, 1993, p. 310).

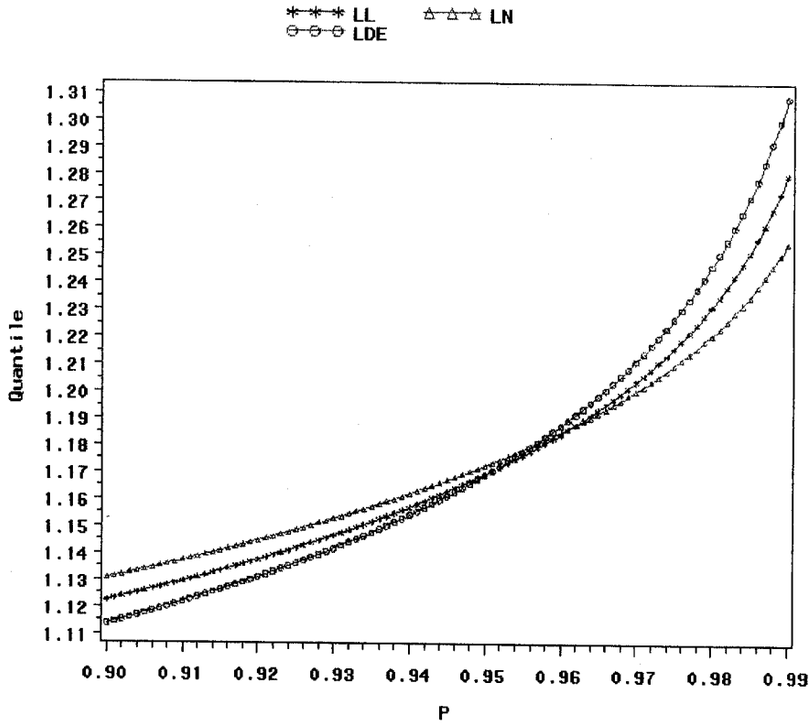

Fig. 1. Quantile Plots of LN, LL, and LDE : (CV = 0.1, $0.90 < p < 0.99$).

Being more revealing for our purposes, the empirical distribution function (EDF) can also be defined in terms of the ordered sample values $y_{(1)} < y_{(2)} \cdots < y_{(n)}$ as

$$F_n(y) = \begin{cases} 0 & \text{if } y < y_{(1)}, \\ \frac{i}{n} & \text{if } y_{(i)} \leqslant y < y_{(i+1)}, \\ 1 & \text{if } y \geqslant y_{(n)}. \end{cases}$$

Direct inversion of the CDF is not very helpful for estimating the quantile function. As $F_{(n)}(y)$ is a step function, its inverse $F_n^{-1}(p)$ exists only for a finite number of values, $p = i/n$, $i = 1, \ldots, n$. Inversion of the EDF does not produce any estimate of $F^{-1}(p)$ if p is not in the set $1/n, 2/n, \ldots, 1$. It is necessary to use interpolation to get estimates of $F(p)$ for all values of p. For example, a linearly interpolated empirical quantile function has the form

$$Q_n(p) = y_{(i)} + n(p - i/n)(y_{(i+1)} - y_{(i)}),$$

where $Q_n(0) = 0$, $Q_n(1) = y_{(n)}$, and i is such that $i/n \leqslant p < (i+1)/n$.

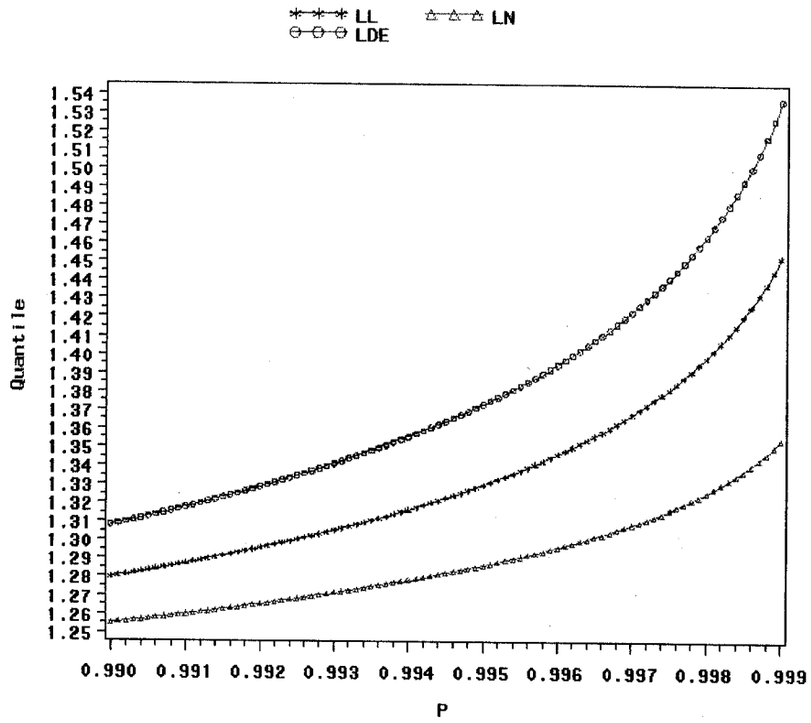

Fig. 2. Quantile Plots of LN, LL, and LDE : (CV = 0.1, $0.99 < p < 0.999$).

According to the EDF all the probability is concentrated between the minimum and maximum of the sample observations. Thus, when estimating a true distribution by an EDF we truncate the tails of the distribution, which alters the resulting estimates of upper and extreme quantiles. In many cases the distribution of the observations in the upper tail is well approximated by a two parameter exponential distribution. Motivated by this fact, Breiman et al. (1979) proposed the tail exponential method for estimating the upper quantiles. As discussed in Ott (1995), this method seeks to fit an exponential distribution to the upper r percent of the observations. Let $y_c = Q_n(1-r)$, where r is a specified tail proportion, usually 10–20 percent. Assume that the conditional distribution of Y given $Y > y_c$ is two parameter exponential with parameters y_c and η. Then, for $y > y_c$, the unconditional distribution function of Y is $F(y) = 1 - r\exp[-(x - x_c)/\eta]$, and hence for $p > 1-r$ the tail exponential quantile function is $Q_n(p) = y_c - \eta \ln((1-p)/r)$. The parameter η of the tail exponential model is estimated as $\hat{\eta} = \bar{y}_c - y_c$ where $\bar{y}_c$ is the mean of the observations above y_c.

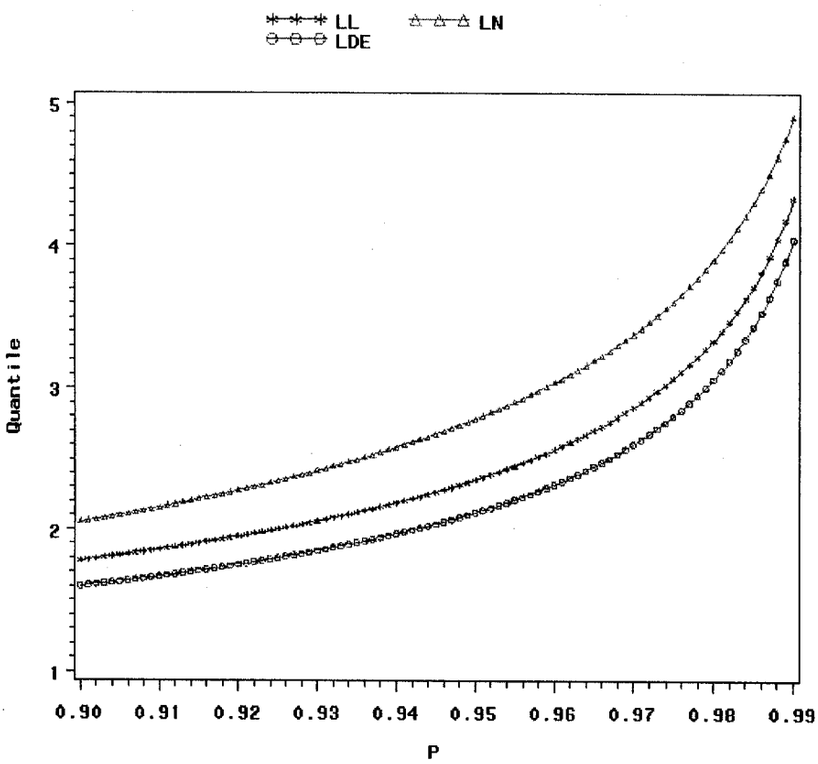

Fig. 3. Quantile Plots of LN, LL, and LDE : (CV = 1.0, $0.90 < p < 0.99$).

4. Simulation study

We performed a simulation study to assess the accuracy of upper and extreme upper quantile estimates for log-normal, log-logistic and log-double exponential distributions. For each model we considered the combinations of $\mu_y = 1$, $\mathrm{CV} = [0.1, 1.0]$, $n = [10, 30, 100, 1000]$, and $p = [0.95, 0.99, 0.999]$. In each case we generated 2000 samples of size n and estimated the pth percentile. Three of the estimates are the maximum likelihood estimates under the assumptions that the population distribution is (i) log-normal, (ii) log-logistic, and (iii) log-double exponential, respectively. For the log-normal distribution the log of the likelihood function is $-n(\ln b\sqrt{2\pi}) - \sum_{i=1}^{n} \ln y_i - 1/2b^2 \sum_{i=1}^{n} (\ln y_i - a)^2$. The log of the likelihood function for the log-logistic distribution is $n(\ln(1/b) + a/b) - (1/b + 1)\sum_{i=1}^{n} \ln y_i$ and for the log-double exponential distribution is $-\ln b - a/b + (1/b - 1)\ln y_i - \ln 2$ if $0 \leqslant y \leqslant \exp(a)$ and $-\ln b + a/b - (1/b + 1)\ln y_i - \ln 2$ if $y \geqslant \exp(a)$.

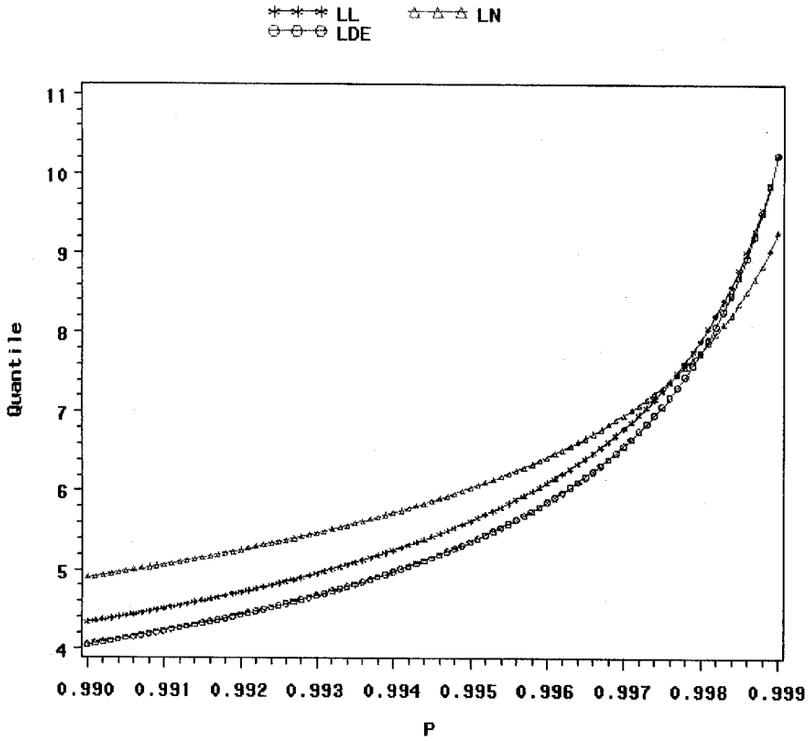

Fig. 4. Quantile Plots of LN, LL, and LDE : (CV = 0.1, $0.99 < p < 0.999$).

The maximum likelihood estimators of the parameters exist in closed form under a log-normal distribution with $\hat{a} = 1/n \ln y_i$ and $\hat{b}^2 = 1/n(\sum_{i=1}^{n} \ln^2 y_i - \hat{a}^2)$ and under a log-double exponential distribution with $\hat{b} = 1/n \sum_{i=0}^{n} |x_i - \hat{a}|$ and $\hat{a} = median(x_i)$. Clearly, in one of these three cases, the model is correctly specified, and in the other two cases the model is mis-specified. We consider the three cases to assess parameter uncertainty as well as effects of incorrect model specification. The fourth estimator is a non-parametric estimator. Table 4 specifies the method of estimation as a function of n and p, i.e., it specifies when we use the empirical quantile function and when we use the tail-exponential method. The fifth estimator is the MLE based on a selected model. In this approach, the selected model is the family, out of log-normal, log-logistic, and log-double exponential, which has the largest maximized likelihood. This fifth estimator is also the MLE under the assumption that the true distribution is either log-normal or log-logistic or log-double exponential.

The simulations were performed using SAS/IML. The inverse transformation method was used to generate variates from the logistic and double exponential distributions.

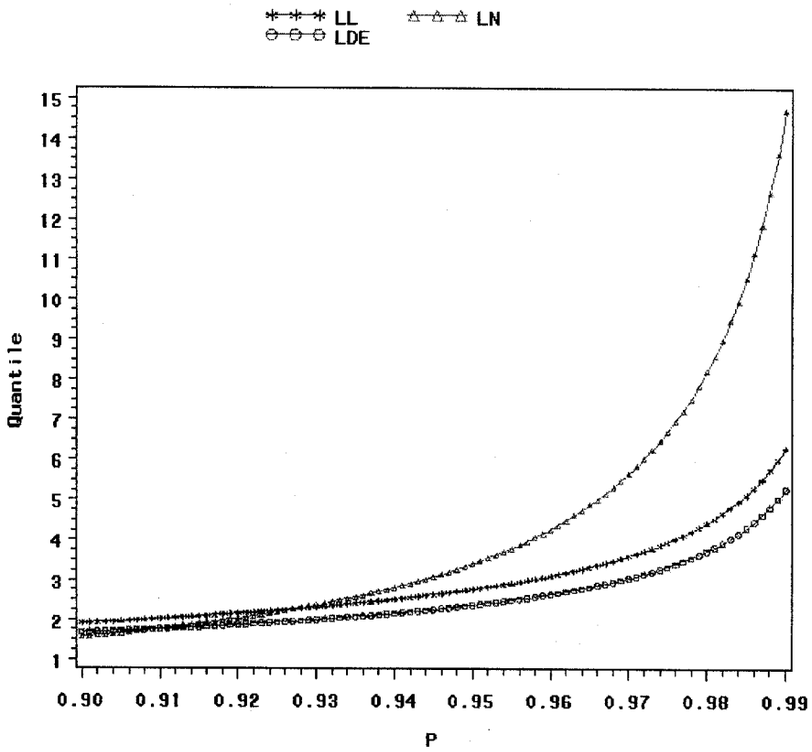

Fig. 5. Quantile Plots of LN, LL, and LDE : (CV = 10.0, $0.90 < p < 0.99$).

Table 4
Methods of estimation used as a function of n and p. The empirical quantile function method is denoted by EQ and TE stands for the tail-exponential method

p			
n	0.95	0.99	0.999
1000	EQ	EQ	EQ
100	EQ	EQ	TE
30	EQ	EQ	TE
10	EQ	TE	TE

Uniform and normal variates were generated using Rannor and Ranuni functions of SAS. The variates were exponentiated to generate values from LL, LN, and LDE distributions. The MLEs of the parameters for log-logistic distribution are obtained by applying the Newton–Raphson root finding procedure to the likelihood equations. We use the criterion of bias and root mean squared error (RMSE) in order to assess the accuracy of the estimates. The simulation results are presented in Tables 5–10.

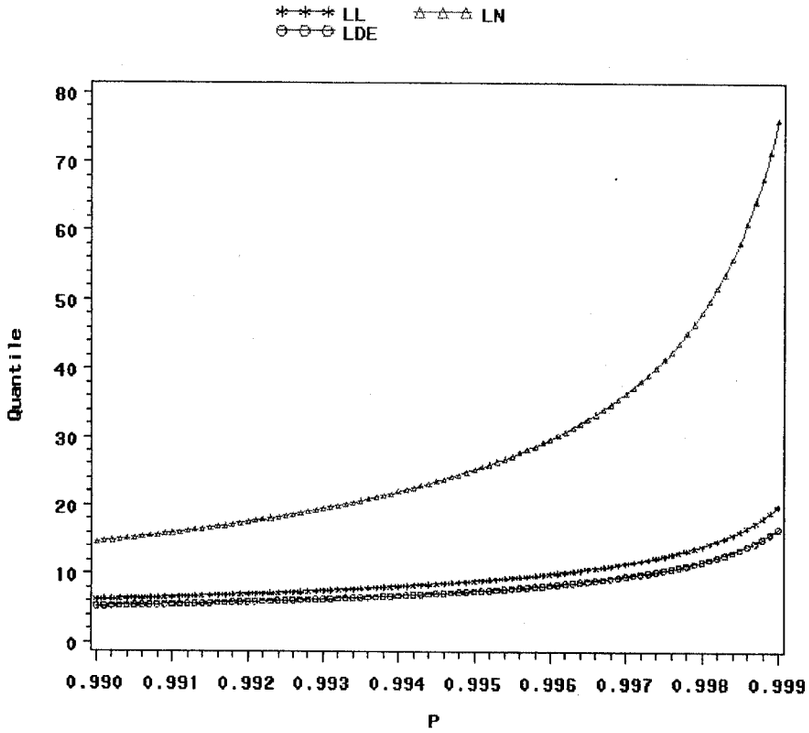

Fig. 6. Quantile Plots of LN, LL, and LDE : (CV = 10.0, $0.99 < p < 0.999$).

The values of bias that are $< 5\%$, $> 25\%$ and 50% of the true parameter values at $p = 0.95$, 0.99 and 0.999 are marked with $*$, $+$, and $>$, respectively.

The first blocks of numbers in Tables 5–10 correspond to the cases where the true model is specified correctly. As expected, the accuracy of the estimators generally increase as the sample size (n) increases and/or the CV decreases. Note that as p increases, i.e., the quantiles become more extreme, the accuracy decreases. The MLEs are not unbiased although the bias decreases to 0 as the sample size increases. The direction and magnitude of bias depends on the specific combination. For small CV (CV = 0.1) the biases are negative but quite small, and the RMSEs are also fairly small. Also, for CV = 0.1, the MLE is most reliable, judged by both bias and RMSE, for the log-normal model, and least reliable for log-double exponential model. For larger CV, however, the biases are generally positive, and they increase in magnitude as n decreases or the CV increases or p increases. Simulation experiments with larger values of CV (CV = 10) indicate that the biases (and the RMSEs) are severe for log-normal model unless n is large. Thus, when the CV is large and the sample size is small or moderate, the MLE should not be used for estimating

Table 5
Bias and RMSE of quantile estimators when the population distribution is log-normal with CV = 0.1

Parent		Log-normal					
CV = 0.1	p	0.95		0.99		0.999	
Treated as	Q	1.1724		1.2549		1.3542	
	n	Bias	RMSE	Bias	RMSE	Bias	RMSE
Log-normal	1000	0.000*	0.005	0.000*	0.007	−0.000*	0.010
	100	−0.000*	0.017	−0.000*	0.023	0.000*	0.032
	30	−0.002*	0.032	−0.000*	0.044	−0.000*	0.058
	10	−0.001*	0.056	−0.005*	0.080	−0.009*	0.107
Log-logistic	1000	0.004*	0.007	0.037*	0.038	0.120	0.121
	100	0.002*	0.019	0.037*	0.046	0.119	0.127
	30	0.000*	0.035	0.032*	0.060	0.111	0.136
	10	−0.006*	0.056	0.022*	0.092	0.091	0.165
Log-DE	1000	0.022*	0.023	0.103	0.103	0.277	0.278
	100	0.022*	0.031	0.098	0.105	0.270	0.277
	30	0.017*	0.043	0.098	0.117	0.259	0.284
	10	0.009*	0.066	0.077	0.135	0.233	0.305
EDF	1000	−0.000*	0.007	−0.002*	0.014	−0.017*	0.037
	100	−0.005*	0.023	−0.021*	0.043	0.033*	0.082
	30	−0.003*	0.042	−0.032*	0.073	0.031*	0.140
	10	−0.008*	0.069	−0.012*	0.129	0.027*	0.246
Selected	1000	0.000*	0.005	0.000*	0.005	0.000*	0.015
	100	0.000*	0.018	0.004*	0.028	0.018*	0.058
	30	−0.001*	0.032	0.010*	0.052	0.034*	0.102
	10	−0.006*	0.061	0.003*	0.087	0.040*	0.151

*Values of bias that are $< 5\%$, $> 25\%$ and 50% of the true value are marked with $*$, $+$, and $>$, respectively.

the upper and extreme quantiles. One should use other estimators with no or smaller bias.

We next consider the second and third blocks in Tables 5–10 to discuss the effect of model mis-specification. The results for $n = 1000$ exhibit the systematic effect of mis-specification as the sampling variability in this case is rather small. When the true distribution is log-normal, from Tables 5 and 6 we see that modeling the data using a log-logistic or log-double exponential distribution usually results in a positive bias. Interestingly, for p=0.95 and $n \leqslant 100$, the bias and RMSE values under the log-logistic assumption are similar to the corresponding values under the correct assumption of log-normality. Thus, mis-specification of a log-normal model by a log-logistic does not seem to have serious consequences for estimating the 95th percentile, unless the sample size is very large. For larger values of p, this mis-specification substantially reduces the accuracy of the estimates. Further, the bias and RMSE increase as p and/or the CV increase. When the data follow a log-normal model, the effects of wrongly using

Table 6
Bias and RMSE of quantile estimators when the population distribution is log-normal with CV = 1

Parent		Log-normal					
CV = 1.0	p	0.95		0.99		0.999	
Treated as	Q	2.7811		4.9049		9.2647	
	n	Bias	RMSE	Bias	RMSE	Bias	RMSE
Log-normal	1000	0.003*	0.108	0.013*	0.243	0.024*	0.59
	100	0.019*	0.367	0.037*	0.814	0.080*	1.91
	30	0.053*	0.687	0.108*	1.537	0.49	3.87
	10	0.106*	1.273	0.448	3.292	1.56	9.21
Log-logistic	1000	0.094*	0.157	1.397+	1.444	9.704 >	9.833
	100	0.096*	0.395	1.376+	1.790	9.796 >	11.02
	30	0.052*	0.701	1.644+	2.618	10.54 >	14.38
	10	0.079*	1.293	1.556+	4.678	12.68 >	29.79
Log-DE	1000	0.480	0.505	4.650 >	4.702	34.69 >	34.99
	100	0.477	0.710	4.619 >	5.104	35.43 >	38.63
	30	0.443	1.019	4.676 >	6.194	36.80 >	48.44
	10	0.547	1.946	5.436 >	10.41	46.91 >	102.4
EDF	1000	−0.005*	0.149	−0.052*	0.461	−0.84	2.03
	100	−0.061*	0.476	−0.534*	1.382	−1.81	2.83
	30	0.066*	0.975	−0.244*	2.081	−1.75	4.04
	10	0.104*	1.770	−0.503*	3.698	−2.21	7.64
Selected	1000	0.000*	0.113	0.010*	0.262	0.080*	0.952
	100	0.014*	0.355	0.248*	1.024	1.92	5.623
	30	0.032*	0.650	0.494	2.021	4.36+	12.486
	10	0.069*	1.261	1.018	4.148	7.88 >	30.10

*Values of bias that are $< 5\%$, $> 25\%$ and 50% of the true value are marked with $*, +$, and $>$, respectively.

log-double exponential distribution are generally much more serious than using the log-logistic to fit the data.

The bias due to mis-specification of a log-logistic distribution by log-normal, reported in Tables 7 and 8, depends on the values of p and CV. When the CV = 0.1, they are generally negative, but when the CV = 1.0, they are negative for $p = 0.99$, 0.999 and positive for $p = 0.95$. For $n = 1000$ these biases are generally quite large, and the RMSEs are much larger than the RMSEs of the MLE under the correct assumption of log-logistic distribution. However, for smaller sample sizes the RMSE values are fairly comparable. Interestingly, the RMSEs under the log-normal assumption are smaller than the RMSEs under log-logistic assumption for $n = 10, 30$, $p = 0.99$, 0.999, and CV = 1.0. Thus, assuming log-normality when the true distribution is log-logistic does not appear to have a major impact on the estimates of the upper and extreme quantiles in samples of size $n \leqslant 30$. In contrast, the assumption of log-double exponential generally induces positive biases, whose magnitude tend to increase with CV

Table 7
Bias and RMSE of quantile estimators when the population distribution is log-logistic with CV = 0.1

Parent		Log-logistic					
CV = 0.1	p	0.95		0.99		0.999	
Treated as	Q	1.1693		1.2800		1.4529	
	n	Bias	RMSE	Bias	RMSE	Bias	RMSE
Log-logistic	1000	−0.000*	0.006	−0.000*	0.009	−0.000*	0.015
	100	−0.000*	0.018	−0.001*	0.028	−0.001*	0.048
	30	−0.003*	0.035	−0.004*	0.053	−0.007*	0.088
	10	−0.010*	0.062	−0.016*	0.094	−0.027*	0.148
Log-normal	1000	0.002*	0.006	−0.025*	0.027	−0.099	0.100
	100	0.000*	0.020	−0.027*	0.039	−0.100	0.108
	30	−0.001*	0.037	−0.030*	0.060	−0.100	0.124
	10	−0.002*	0.064	−0.034*	0.096	−0.105	0.167
Log-DE	1000	0.015*	0.017	0.059*	0.060	0.142	0.143
	100	0.014*	0.026	0.056*	0.067	0.140	0.154
	30	0.010*	0.040	0.051*	0.083	0.129	0.176
	10	0.005*	0.065	0.035*	0.117	0.103	0.234
EDF	1000	−0.000*	0.009	−0.003*	0.021	0.013*	0.070
	100	−0.005*	0.028	−0.027*	0.065	−0.013*	0.106
	30	0.000*	0.053	−0.006*	0.100	−0.023*	0.198
	10	−0.028*	0.075	−0.031*	0.164	−0.027*	0.336
Selected	1000	−0.000*	0.005	−0.000*	0.010	−0.000*	0.018
	100	0.000*	0.020	−0.003*	0.039	−0.016*	0.095
	30	−0.001*	0.037	−0.0102*	0.066	−0.030*	0.131
	10	−0.011*	0.062	−0.020*	0.098	−0.055*	0.181

*Values of bias that are $< 5\%$, $> 25\%$ and 50% of the true value are marked with $*$, $+$, and $>$, respectively.

and p. Their overall properties are markedly inferior except possibly for CV = 0.1 and $p = 0.95$.

Tables 9 and 10 report the effects of mis-specification of a true log-double exponential distribution. The MLEs under log-logistic assumption are negatively biased, but their RMSEs are close to the RMSEs under log-double exponential for small-to-moderate sample sizes. Actually, for $n \leqslant 30$, the RMSEs under mis-specification by log-logistic are mostly smaller than the RMSEs under the correct assumption of log-double exponential distribution. Mis-specification by the log-normal leads to negative biases for $p = 0.99$ and 0.999; for $p = 0.95$, they are positive for CV = 1. The RMSEs are close (and even smaller in several cases) to those under the correct assumption for smaller ($n \leqslant 30$) sample sizes and $p \leqslant 0.99$. When $p = 0.999$, however, the negative bias remained and the RMSEs were high regardless of which model was fit. The log-normal mis-specification is slightly worse than the log-logistic mis-specification. Overall, mis-specification of log-double exponential by log-normal or log-logistic is

Table 8
Bias and RMSE of quantile estimators when the population distribution is log-logistic with CV = 1

Parent		Log-logistic					
CV = 1.0	p	0.95		0.99		0.999	
Treated as	Q	2.3507		4.3369		10.224	
	n	Bias	RMSE	Bias	RMSE	Bias	RMSE
Log-Logistic	1000	0.003*	0.081	0.007*	0.221	0.003*	0.71
	100	0.007*	0.263	0.025*	0.719	0.079*	2.34
	30	0.014*	0.494	0.015*	1.296	0.424*	4.83
	10	−0.021*	0.910	0.104*	2.665	1.63	14.8
Log-normal	1000	0.038*	0.097	−0.569	0.597	−3.900+	3.92
	100	0.033*	0.294	−0.529	0.803	−3.833+	4.04
	30	0.062*	0.541	−0.416	1.271	−3.070+	4.511
	10	0.093*	0.989	−0.227	2.656	−2.814+	8.076
Log-DE	1000	0.231	0.253	1.574	1.612	9.092 >	9.258
	100	0.220	0.395	1.542	1.885	9.286 >	11.01
	30	0.234	0.649	1.597	2.635	10.30 >	16.26
	10	0.257	1.163	1.690	4.853	12.54 >	34.27
EDF	1000	−0.002*	0.128	−0.043*	0.519	−1.04	3.34
	100	−0.053*	0.444	−0.418	1.416	−3.39+	5.22
	30	0.100*	0.953	−0.065*	3.395	−3.33+	6.43
	10	−0.062*	1.510	−0.442	4.308	−4.06+	8.60
Selected	1000	−0.000*	0.085	−0.004*	0.224	−0.022*	0.870
	100	0.027*	0.294	−0.002*	1.026	−0.053*	4.847
	30	0.006*	0.524	−0.060*	1.611	0.301*	8.510
	10	0.003*	0.958	0.040*	2.808	0.576	17.06

*Values of bias that are $< 5\%$, $> 25\%$ and 50% of the true value are marked with $*$, $+$, and $>$, respectively.

fairly innocuous for estimating upper percentiles from small or moderate sized data sets.

Considering effects of all incorrect specifications of the model, we conclude that when the sample size is not very large, the assumption of log-normality is fairly harmless for estimating upper percentiles. Extreme upper percentiles are difficult to estimate accurately even when one knows the correct underlying model. When the sample size is large, significant biases result from assuming an incorrect model. In such cases, however, it is much easier to identify the correct model using goodness of fit tests or the selected estimator can be used.

Now we discuss the behavior of the EDF estimator. Expectedly, the RMSE values are fairly large in the cases where the fully non-parametric method has been used (see Table 4). They are mostly larger than the RMSE values under the log-normal assumption, even when that assumption is not correct. However, the tail exponential approach performs fairly well for log-normal and log-logistic distributions, especially

Table 9
Bias and RMSE of quantile estimators when the population distribution is log-double exponential with CV=0.1

Parent		Log double-exponential					
CV = 0.1	p	0.95		0.99		0.999	
Treated as	Q	1.1689		1.3082		1.5367	
	n	Bias	RMSE	Bias	RMSE	Bias	RMSE
Log-DE	1000	−0.000*	0.006	−0.000*	0.011	0.000*	0.021
	100	−0.000*	0.021	−0.000*	0.037	−0.001*	0.069
	30	−0.002*	0.038	−0.004*	0.067	−0.005*	0.118
	10	−0.007*	0.068	−0.013*	0.118	−0.015*	0.221
Log-logistic	1000	−0.011*	0.012	−0.047*	0.048	−0.116	0.117
	100	−0.011*	0.023	−0.047*	0.057	−0.116	0.128
	30	−0.011*	0.037	−0.051*	0.077	−0.127	0.158
	10	−0.018*	0.066	−0.064	0.116	−0.131	0.211
Log-normal	1000	0.001*	0.007	−0.055*	0.056	−0.185	0.186
	100	0.000*	0.0237	−0.056*	0.067	−0.186	0.192
	30	−0.000*	0.043	−0.059	0.084	−0.192	0.211
	10	−0.005*	0.072	−0.068	0.126	−0.202	0.250
EDF	1000	−0.000*	0.011	−0.004*	0.028	−0.043*	0.097
	100	−0.006*	0.035	−0.034*	0.080	−0.040*	0.142
	30	0.002*	0.065	−0.014*	0.123	−0.045*	0.252
	10	−0.021*	0.092	−0.031*	0.224	−0.059*	0.437
Selected	1000	0.000*	0.006	−0.000*	0.011	0.000*	0.021
	100	−0.003*	0.021	−0.005*	0.0245	−0.026*	0.088
	30	−0.007*	0.039	−0.032*	0.0786	−0.073	0.160
	10	−0.017*	0.068	−0.051*	0.126	−0.122	0.246

*Values of bias that are $< 5\%$, $> 25\%$ and 50% of the true value are marked with $*$, $+$, and $>$, respectively.

for large CV. But, it does not work well for log-double exponential distribution. In that case, even the MLE under the incorrect assumption of log-normality performs better.

Finally, we discuss performance of the selection estimator, which is the MLE under the selected model. The selection probabilities of different families, for different true distributions with CV = 1, are presented in Table 11. We report them only for CV = 1 as they changed very little with the CV. For $n = 1000$, the true model is selected with probability at least 0.98, and hence the resulting estimator generally performs almost as well as the MLE under the correct distribution. Only for log-normal data with CV=1.0, the estimators of 99th, and 99.9th percentiles have noticeably larger RMSE than the MLE for the log-normal. For $n \leqslant 100$, its performance is roughly comparable to the MLE under the log-normal assumption; neither one is uniformly better than the other.

Based on the simulation results we come to the following conclusions. For large n, we suggest using the data to select a model, and then estimate the quantiles based on the selected model. For smaller sample sizes it is difficult to identify the correct

Table 10
Bias and RMSE of quantile estimators when the population distribution is log-double exponential with CV=1

Parent		Log double-exponential					
CV = 1.0	p	0.95		0.99		0.999	
Treated as	Q	2.1180		4.0504		10.2410	
	n	Bias	RMSE	Bias	RMSE	Bias	RMSE
Log-DE	1000	−0.001*	0.068	−0.000*	0.20	0.018*	0.81
	100	−0.006*	0.218	−0.001*	0.66	0.222*	2.74
	30	−0.008*	0.418	0.079*	1.42	0.761	6.23
	10	−0.000*	0.781	0.250	2.87	2.80+	16.7
Log-logistic	1000	−0.112	0.130	−0.775	0.790	−3.73+	3.76
	100	−0.116	0.234	−0.773	0.911	−3.63+	3.90
	30	−0.093*	0.393	−0.702	1.194	−3.46+	4.57
	10	−0.086*	0.7301	−0.691	2.042	−2.81+	7.80
Log-normal	1000	0.020*	0.081	−0.890	0.905	−5.36 >	5.37
	100	0.017*	0.255	−0.889	1.023	−5.32 >	5.423
	30	0.035*	0.504	−0.884	1.297	−5.22 >	5.642
	10	0.062*	0.993	−0.708	2.729	−4.77+	7.138
EDF	1000	−0.005*	0.11	−0.059*	0.52	−1.01	3.93
	100	−0.066*	0.36	−0.380	1.58	−3.79+	4.79
	30	0.090*	0.79	−0.222	2.71	−3.83+	7.07
	10	−0.041*	1.36	−0.445	4.16	−4.33+	9.18
Selected	1000	0.000*	0.065	0.000*	0.213	0.001*	0.847
	100	−0.027*	0.224	−0.159*	0.739	−0.468*	3.133
	30	−0.046*	0.407	−0.336	1.321	−1.136	5.950
	10	−0.062*	0.755	−0.465	2.403	−2.153	11.75

*Values of bias that are $< 5\%$, $> 25\%$ and 50% of the true value are marked with $*$, $+$, and $>$, respectively.

Table 11
Selection probabilities for the parent distribution with CV = 1

Parent	LL			LN			LDE		
Selected	LL	LN	LDE	LL	LN	LDE	LL	LN	LDE
n									
1000	0.99	0.01	0.00	0.01	0.99	0.00	0.00	0.00	1.00
100	0.53	0.31	0.16	0.16	0.83	0.11	0.13	0.02	0.85
30	0.23	0.50	0.27	0.12	0.76	0.12	0.16	0.18	0.66
10	0.07	0.60	0.33	0.06	0.70	0.24	0.07	0.43	0.50

model (see Table 11 and Haas, 1997), and hence one cannot rely on the estimates derived under the selected model. In such cases, one should try to choose a model by examining the subject matter and related studies. If one has considerable uncertainty about the correct model, we believe one should obtain multiple estimates using different

Table 12
Nickel concentrations (ppb) at four monitoring wells, USEPA (1992)

Observations ($n = 20$)						
58.8	1.0	262.0	56.0	8.7	19.0	81.5
331.0	14.0	64.4	39.0	151.0	27.0	21.4
578.0	3.1	942.0	85.6	10.0	637.0	—

Summary statistics						
Min	Max	Mean	SD	Med.	Skew.	Kurt.
1.0	942.0	169.52	259.7	57.4	2.0	3.41

methods. For moderate sample sizes, the MLE under the log-normal assumption, and the selection estimates appear to be two reasonable alternatives. For small sample sizes, selection estimates are unreliable, especially for large p. For estimating upper and extreme quantiles based on a small sample ($n \leqslant 30$), we suggest reporting the tail exponential estimates, and the MLE under the log-normal assumption. The inherent model uncertainty should be reflected in the differences between them.

5. An example

Nickel is a metal found only in combined form in nature. Used in electronics industry, coal gasification, petroleum refining, and hydrogenation of fats and oils, nickel is a potential ground and surface water pollutant. There is currently no legal limit on the amount of nickel in drinking water. Nickel has not been found to potentially cause health effects from acute exposures at levels below 0.1 mg/l. USEPA (2001) contains various fact sheets about this and other contaminants. We will investigate estimation of the upper quantiles of a data set of nickel concentrations from four monitoring wells. The data set appears in a guidance document on analysis of ground-water monitoring data (USEPA, 1992) and is discussed by Millard (1998).

The data set consists of $n=20$ nickel concentrations in parts per billion. We use these data, which appear in Table 12 along with some summary measures, for illustrative purposes. Millard (1998) considered $\mathscr{F}$ to only contain the log-normal distribution. We will enlarge $\mathscr{F}$ to include log-logistic and log-double exponential families. Table 13 displays the values of the MLE, log likelihood, and three estimated quantiles for all three distributions. The model with the largest likelihood is log-normal. Note that none of the listed distributions are rejected at a 05% confidence level. The log of the likelihoods are very close. Anderson–Darling test for normality of ln(nickel) produces a p-value in excess of 0.25. The same test for a logistic and double exponential distributions produce p-values 0.15. There seems to be some uncertainty about the underlying model.

Estimated quantiles based on the tail-exponential method corresponding to a 10, and a 20 percent tail proportion also appear in Table 13. The estimates vary substantially for each value of p. The exponential method leads to estimated quantiles that are smaller than the ones obtained under parametric models. As there is considerable uncertainty

Table 13
Summary of fitting distributions to nickel concentrations and the estimated upper quantiles

Distribution	Parameter	MLE	Log likelihood	Estimated quantiles		
				$p = 0.95$	$p = 0.99$	$p = 0.999$
L-logistic	(a, b)	(3.94, 1.01)	−118.37	1032.7	5527.1	57905.7
L-normal	(μ, σ)	(3.91, 1.75)	−118.01	903.7	2990.2	11433.8
Log-DE	(a, b)	(4.05, 1.39)	−118.9	1426.9	13487.2	335396.1
TE method	$r = 0.20$			761.0	1340.4	2169.3
TE method	$r = 0.10$			724.60	1064.9	1551.9

about an effective model, we believe one should utilize the estimates under different models to draw conclusions. Estimates under the log-double exponential model are substantially larger than all other estimates. The estimates under the log-normal and log-logistic distributions are comparable for $p = 0.95$, but differ substantially for $p = 0.999$. In this example, one might consider an interval at the 95th(99th) percentile ranging from the log-normal estimates (903.7)(2990.2) to (1032.2)(5527.1) from the log-logistic model. A formal confidence interval would be much larger.

6. Summary and recommendations

In this paper, we considered the accuracy of upper and extreme tail estimates of three right skewed distributions under model and parameter uncertainty. We used the criteria of bias and root mean squared error in order to assess the accuracy of the estimates. The distributions considered are log-transformation of three well-known and symmetric distributions, the log-normal, log-logistic and log-double exponential distributions. We examined and compared performances of the MLE and non-parametric estimators based on the empirical or a quasi-empirical quantile function (tail-exponential method). We considered four cases that are encountered in practice. In particular, we considered the cases where (i) the model is correctly specified, (ii) the model is mis-specified, (iii) the best model is selected using the data, and (iv) no form is assumed for the model.

In practice it is important to report standard errors or confidence intervals along with the point estimates to provide information about the reliability of the point estimates. Thus, the effects of model mis-specification on the width and coverage probability of confidence intervals deserve further investigation. For a given dataset, the true (correct) model is unknown and its analysis should be guided by (i) identification of a reasonable model and (ii) the robustness of the proposed methodology. We should know how an analysis based on an assumed model perform when it differs from the true model. Identification of a useful and effective model is easier for large datasets. For smaller datasets, one needs to rely on more robust methods. Fully non-parametric methods are robust, but may not be very efficient for some models. It is useful, therefore, to use a method that works well for a class of sufficiently realistic models. In this paper we have studied the robustness of estimates of upper and extreme percentiles when the

true model is either log-normal, log-logistic or log-double exponential, all of which are symmetric location-scale families on the log scale.

Generally speaking, when the model is specified correctly, the accuracy of the estimators increase as the sample size increases and/or the CV decreases and/or p decreases. Under model mis-specification, we observed the following. The assumption of log-normality when the true distribution is log-logistic does not appear to have a major effect on the estimates of upper and extreme quantiles. The mis-specification of a log-double exponential by a log-normal or log-logistic is fairly innocuous for estimating upper percentiles from small or moderate sized data sets but become quite noticeable for larger samples. Considering effects of all the incorrect specifications of a model, we conclude that when the sample size is not very large, the assumption of log-normality is relatively harmless for estimating the upper percentiles. The extreme percentiles were difficult to reliably estimate in modest sized samples for all three distributions. When the sample size is large, significant biases result from assuming an incorrect model. In such cases, however, it is much easier to identify the correct model using goodness of fit tests.

When the size of the sample is large we should use the data to select a model, and then estimate the quantiles based on the selected model. For smaller sample sizes it is difficult to identify the correct model and hence one cannot rely on the estimates derived under the selected model. In such cases, one should try to choose a model by examining the subject matter and related studies. But, if one has considerable uncertainty about the correct model, we believe it is helpful to obtain multiple estimates using different methods. For moderate sample sizes, the MLE under log-normal assumption, and the selection estimates appear to be two reasonable alternatives. Considering the non-parametric estimators, the tail exponential approach works fairly well for log-normal and log-logistic distributions, especially when the CV is large. Unfortunately, it does not work well for log-double exponential distribution. For small sample sizes, selection estimates are unreliable, especially for large p. For estimating upper and extreme quantiles based on a small sample ($n \leqslant 30$), we suggest reporting the tail exponential estimates, and the MLE under log-normal assumption.

The coefficient of variation has a great impact on the results in all four situations. Even when the parent distribution is identified a priori, the extreme upper tail estimates are not accurate at small sample sizes and large values of CV. Caution must be exercised when identifying a distribution a priori as model mis-specification can result in high bias and mean squared error at large values of CV. Large values of CV also impact the identification of the correct model. Estimates based on the selection method may be suspect for small sample sizes and high values of CV. Large sample sizes are necessary to reduce the mis-classification rates.

Acknowledgements

We would like to express our appreciation to a referee whose comments improved the presentation of this article.

References

Aitchison, J., Brown, J.A.C., 1973. The Lognormal Distribution. Cambridge Press, Cambridge.

Aitchison, A.C., 1982. Regression diagnostics, transformations and constructed variables. J. Roy. Statist. Soc. Ser. B 1, 1–36.

Andrews, D.F., et al., 1972. Robust Estimation of Location. Princeton University Press, Princeton.

Bennett, S., 1983. Log-logistic regression models for survival data. Appl. Statist. 32 (2), 165–171.

Breiman, L., Gins, J., Stone, C., 1979. New Methods for Estimating Tail Probabilities and Extreme Value Distributions. Technology Service Corp., Santa Monica, CA, TSC-PD-A2261.

Burmaster, D.E., 1998. Lognormal distribution for skin area as a function of body weight. Risk Anal. 18 (1), 27–32.

Burmaster, D.E., Crouch, A.C., 1997. Lognormal distributions for body weight as a function of age for males and females in the United States, 1976–1980. Risk Anal. 17 (4), 499–505.

D'Agostino, R.B., Stephens, M.A., 1986. Goodness-of-Fit Techniques. Marcel Dekker, New York.

Draper, D., 1995. Assessment and propagation of model uncertainty. J. R. Statist. Soc. Ser. B 57 (1), 45–97.

Dubey, S.D., 1966. Transformation for estimation of parameters. J. Ind. Statist. Assoc. 4, 109–124.

Dumonceaux, R., Antle, C.E., 1973. Likelihood ratio test for discrimination between two models with unknown location and scale parameters. Technometrics 15 (1), 19–27.

Dyer, A.R., 1973. Discrimination procedures for separate families of hypotheses. J. Amer. Statist. Assoc. 68, 970–974.

Efron, B., Tibshirani, R.J., 1993. An Introduction to Bootstrap. Chapman & Hall, London.

Frey, H.C., Burmaster, D.E., 1999. Methods for characterizing variability and uncertainty: comparison of bootstrap and likelihood-based approaches. Risk Anal. 19 (1), 109–130.

Firth, D., 1988. Multiplicative errors: log-normal or gamma?. J. R. Statist. Soc. B 50 (2), 266–268.

Gastwirth, J.L., 1966. On robust procedures. J. Amer. Statist. Assoc. 61 (316), 929–948.

Haas, C.N., 1997. Importance of distributional form in characterizing inputs to Monte Carlo risk assessment. Risk Anal. 17 (1), 107–113.

Hoaglin, D.C., Mosteller, F., Tukey, J.W. (Eds.), 1983. Understanding Robust and Exploratory Data Analysis. Wiley, New York.

Johnson, N.L., 1949. Systems of frequency curves generated by methods of translation. Biometrika 36, 149–176.

Johnson, N.L., Kotz, S., Balakrishnan, N., 1995. Continuous Univariate distributions, Volumes I and II, 2nd Edition. Wiley, New York.

Kappenman, R.F., 1982. On a method for selecting a distributional model. Commun. Statist. Theory Methods 11 (6), 663–672.

Kapteyn, J.C., 1903. Skew Frequency Curves in Biology and Statistics. Astronomical Laboratory. Groningen, Noordhoff.

Lehman, E.L., 1983. Theory of Point Estimation. Wiley, New York.

Mage, D.T., 1981. A Review of Applications of Probability Models for Describing Aerometric Data. Environmetrics 81: Selected Papers. SIAM, Philadelphia, PA.

McCullagh, P., Nelder, J.A., 1989. Generalized Linear Models. Chapman & Hall, London.

Millard, S.P., 1998. Environmental Stats for S-Plus. Springer, Berlin.

Mudholkar, G.S., George, E.O., 1978. A remark on the shape of the logistic distribution. Biometrika 65 (3), 667–669.

Murray, D.M., Burmaster, D.E., 1994. Estimated distributions for average daily consumption of total and self-caught fish for adult Michigan angler households. Risk Anal. 14 (4), 513–519.

Ott, W.R., 1995. Environmental Statistics and Data Analysis. Lewis Publishers, Boca Raton, Florida.

Quesenberry, C.P., Kent, J., 1982. Selecting among probability distributions used in reliability. Technometrics 24 (1), 59–66.

Rosebury, A.M., Burmaster, D.E., 1992. Log-normal distributions for water intake for children and adults. Risk Anal. 12 (1), 99–104.

Rustagi, J.S., 1964. Stochastic behavior of trace substances. Arch. Environ. Health 8, 84–87.

Séménou, M., 1996. Quantile estimation under possibly misspecified generalized linear model. Statist. Probab. Lett. 27, 357–365.

Tadikamalla, P.R., Johnson, N.L., 1982. Systems of frequency curves generated by transformations of logistic variables. Biometrika 69 (2), 461–465.

Uppuluri, V.R.R., 1980. Some properties of log-laplace distribution. Manuscript No. 103, International Summer School on Statistical Distributions in Scientific Work Trieste, Italy, July-August.

Wiens, B.L., 1999. When log-normal and gamma models give different results: a case study. Amer. Statist. 53 (2), 89–93.

USEPA, 1985. Technical Support Documents for Water Quality-Based Toxics Control. Washington DC.

USEPA, 1992. Statistical analysis of ground-water monitoring data at RCRA facilities. Office of Solid Waste, Washington, DC, EPA/530-R-93-003.

USEPA, 2001. Web site of the Office of Water at the United States Environmental Protection Agency: (www.epa.gov/safewater).